(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平10－312246

(43)公開日　平成10年(1998)11月24日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ０６Ｆ	3/06	３０５	Ｇ０６Ｆ	3/06	３０５Ｃ
		５４０			５４０
	12/16	３２０		12/16	３２０Ｌ
Ｇ１１Ｂ	20/18	５７０	Ｇ１１Ｂ	20/18	５７０Ｚ

審査請求　未請求　請求項の数４　ＯＬ　（全 14 頁）

(21)出願番号　特願平9－120487

(22)出願日　平成９年(1997)５月12日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　山本　康友
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　山本　彰
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　佐藤　孝夫
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内

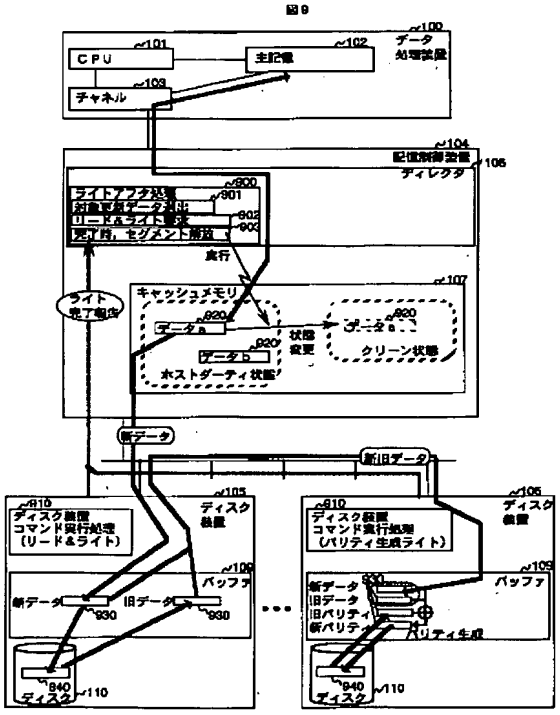
(74)代理人　弁理士　小川　勝男

(54)【発明の名称】　記憶装置サブシステム

(57)【要約】

【課題】更新データに対する更新パリティ生成における負荷を軽減し、さらに更新前パリティリードや更新データライトにおける平均シーク／サーチ時間を短縮することで、記憶装置サブシステム全体のスループット向上を図る。

【解決手段】更新データのパリティ生成およびライトを行う際に、パリティグループ内の一つ以上の更新データの新データを各データディスク装置１０５に転送して、各データディスク装置１０５では、旧データをリード後、新データをライトし、新旧データをパリティディスク装置１０５に転送する。パリティディスク装置１０５では旧パリティリードと新パリティ生成およびライトを実行する。パリティ生成のための排他的論理和演算をパリティディスク装置１０５で一括実行することで、演算に要するオーバヘッドを削減する。



1

【特許請求の範囲】

【請求項１】それぞれのパリティグループがｍ個（ｍ≧１）のデータと前記データを回復するための冗長データから構成され、少なくとも１つ以上の前記パリティグループに属する１つ以上の前記データ、あるいは、１つ以上の前記冗長データを格納する記憶媒体を有する１つ以上の記憶装置と、データ処理装置と、前記データ処理装置と前記記憶装置との間に介在し、前記データ処理装置と前記記憶装置とのデータ転送を制御する記憶制御装置からなる記憶装置サブシステムにおいて、前記記憶制御装置が、前記パリティグループに属する一つ以上の前記データの更新値をね前記データを格納する一つ以上の第１の記憶装置に送信する手段を有し、前記第１の記憶装置が前記記憶制御装置から前記データの更新値を受信する手段と、前記データの更新前の値を前記記憶媒体から読み出す手段と、前記データの更新値を前記記憶媒体に書き込む手段と、前記データの更新値と更新前の値を前記冗長データを格納する第２の記憶装置に直接送信する手段を有し、前記第２の記憶装置が前記第１の記憶装置から前記データの更新値と更新前の値を受信する手段と、前記冗長データの更新前の値を前記記憶媒体から読み出す手段と、前記データの更新値と更新前の値と前記冗長データの更新前の値から前記冗長データの更新値を作成する手段と、前記冗長データの更新値を前記記憶媒体に書き込む手段を有することを特徴とする記憶装置サブシステム。

【請求項２】それぞれのパリティグループがｍ個（ｍ≧１）のデータと前記データを回復するための冗長データから構成され、少なくとも１つ以上の前記パリティグループに属する１つ以上の前記データ、あるいは、１つ以上の前記冗長データを格納する記憶媒体を有する１つ以上の記憶装置と、データ処理装置と、前記データ処理装置と前記記憶装置との間に介在し、前記データ処理装置と前記記憶装置とのデータ転送を制御する記憶制御装置からなる記憶装置サブシステムにおいて、前記記憶制御装置が、前記パリティグループに属する一つ以上の前記データの更新値を、前記データを格納する１つ以上の第１の記憶装置に送信する手段と、前記データの更新前の値を前記第１の記憶装置から受信する手段と、前記データの更新値と更新前の値を前記冗長データを格納する第２の記憶装置に送信する手段を有し、前記第１の記憶装置が前記記憶制御装置から前記データの更新値を受信する手段と、前記データの更新前の値を前記記憶媒体から読み出す手段と、前記データの更新値を前記記憶媒体に書き込む手段と、前記データの更新前の値を前記記憶制御装置に送信する手段を有し、前記第２の記憶装置が前記記憶制御装置から前記データの更新値と更新前の値を受信する手段と、前記冗長データの更新前の値を前記記憶媒体から読み出す手段と、前記データの更新値と更新前の値と前記冗長データの更新前の値から前記冗長データの更新値を作成する手段と、前記冗長データの更新値を前記記憶媒体に書き込む手段を有することを特徴とする記憶装置サブシステム。

2

【請求項３】請求項２で、前記記憶制御装置が、複数の前記データの更新値と更新前の値を同時に格納できるキャッシュメモリを有し、前記第１の記憶装置から前記データの更新前の値を受信後、前記データの更新値とともに前記キャッシュメモリに溜め置き、前記データの前記第１の記憶装置への反映処理とは非同期に、シーク／サーチ動作時間を考慮したスケジュールにより、前記第２の記憶装置上の前記冗長データへの反映処理を行うことを特徴とする記憶装置サブシステム。

【請求項４】請求項１で、前記第２の記憶装置が、複数の前記データの更新値と更新前の値を同時に格納できるバッファメモリを有し、前記第１の記憶装置から前記データの更新値と更新前の値を受信後、バッファメモリに溜め置き、前記データの前記第１の記憶装置への反映処理とは非同期に、シーク／サーチ動作時間を考慮したスケジュールにより、前記第２の記憶装置上の前記冗長データへの反映処理を行うことを特徴とする記憶装置サブシステム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、ディスクアレイ向きの高機能ディスク装置、および、ディスクアレイ向きの高機能ディスク装置と記憶制御装置により構成される記憶装置サブシステムに関する。

【０００２】

【従来の技術】発明に最も近い公知例として、シカゴのイリノイ大学で開かれたＡＣＭＳＩＧＭＯＤ　会議において発表されたＤ．Ｐａｔｔｅｒｓｏｎ，Ｇ．ｇｉｂｓｏｎ，ａｎｄ Ｒ．Ｈ．Ｋａｒｔｚ；Ａ Ｃａｓｅ ｆｏｒ Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙｓ ｏｆ Ｉｎｅｘｐｅｎｓｉｖｅ Ｄｉｓｋｓ（ＲＡＩＤ），ＡＣＭ ＳＩＧＭＯＤ Ｃｏｎｆｅｒｅｎｃｅ，Ｃｈｉｃａｇｏ，ＩＬ，（Ｊｕｎｅ １９８８），ｐｐ．１０９－１１６なる論文が知られている。

【０００３】Ｐａｔｔｅｒｓｏｎの論文は、ディスクアレイ上のデータ配置に関する技術を開示したものである。ディスクアレイシステムでは、物理的には複数のディスク装置をデータ処理装置に対しては１台の論理的なディスク装置に見せかけることで高性能化を実現している。また、データを格納したディスク装置に障害が発生した場合、障害が発生したディスク装置に記憶されているデータの回復を行うことを可能とする冗長データをデータを記憶したディスク装置とは別のディスク装置に格納しておくことにより高信頼化を達成している。

【０００４】通常、ディスク装置に格納されたデータのリード／ライト単位をレコードと呼ぶ。ただし、ディスクアレイの場合、データ処理装置から見たリードライト単位であるレコードと、ディスク装置に実際に記録され

るレコードとではデータ長が異なる場合がある。以下、前者を論理レコード、後者を物理レコードと呼ぶ。

【０００５】Pattersonの論文では、いくつかの論理レコードの配置方法がＲＡＩＤ１からＲＡＩＤ５の５段階のレベルのＲＡＩＤ構成として紹介され、性能および信頼性の検討結果が報告されている。このうち、ＲＡＩＤ５のディスクアレイは各論理レコードを１つの物理レコードとしてディスク装置に格納し、互いに異なるディスク装置に格納されたｍ（ｍ≧１）個の論理レコードに対して、その排他的論理和であるパリティと呼ばれる冗長データを作成し、１つの物理レコードとしてディスク装置に格納したものである。以下、処理装置が直接リード／ライトする論理レコードを格納した物理レコードをデータレコード、パリティを格納した物理レコードをパリティレコード、データレコードを記憶するディスク装置をデータディスク装置、パリティレコードを記憶するディスク装置をパリティディスク装置と呼ぶ。また、ｍ個のデータレコードとパリティレコードから構成されるレコードのグループをパリティグループと呼ぶ。ＲＡＩＤ５のディスクアレイは、更新のあったデータレコードに対して、パリティ生成時にデータレコードとパリティレコードの更新前の値である旧データと旧パリティをそれぞれのレコードが格納されるディスク装置からリードし、データレコードの更新値である新データと排他的論理和をとってパリティレコードの更新値である新パリティを生成し、ディスク装置へ新パリティをライトする必要があるため、ライト処理に付随するオーバヘッドは大きい。このようなパリティ生成処理を効率よく行い、システムの高性能化を実現する技術としては次のような技術がある。

【０００６】特開平４－２４５３４２号に開示される技術では、論理レコード（＝データレコード）に対する更新処理をキャッシュ上で実行し、新パリティの作成や新データと新パリティのディスク装置への書き込みは後から実行する。ある更新データレコードに対して新パリティの作成を開始する前に、同じパリティレコードに対応する他のデータレコードに更新があった場合、新パリティの作成をまとめて実行することにより高性能化を実現している。このように、処理装置からの論理レコード更新に対し、ディスク装置への反映を非同期で行う方式を一般にライトアフタと呼ぶ。

【０００７】特開平５－４６３２４号には、新パリティの生成に必要な情報をパリティレコードのあるパリティディスク装置に転送し、ディスク装置側で新パリティを生成することで、記憶制御装置の負荷や、データ更新時に発生する記憶制御装置とディスク装置との間のデータ転送回数を削減し、ディスクアレイシステムの高性能化を図る技術が開示されている。この技術では、パリティ生成に必要な情報をパリティディスク装置に送出する方法の一つとして、記憶制御装置と各ディスク装置をブロードキャスト可能なネットワークで接続し、パリティグループの１データの新データをデータディスク装置とパリティディスク装置にブロードキャストして、旧データはデータディスク装置からパリティディスク装置へ直接転送することで、インタフェースにおける転送量を低減する方法が開示されている。

【０００８】また、転送量の低減を図る技術としては、特開平７－２６１９４６号で、更新データレコードの新データをデータディスク装置に転送し、ディスク装置側で旧データを読んで新データを書き込み、新旧データから中間値を生成し、生成した中間値をパリティディスク装置に直接転送する技術が開示されている。新旧データをそのまま転送するのではなく、中間値に変換して転送することで、データディスク装置とパリティディスク装置間のデータ転送量を低減できる。

【０００９】

【発明が解決しようとする課題】ＲＡＩＤ５のディスクアレイを用いた場合、記憶制御装置がデータ処理装置からライト要求を受けて、データレコードを更新するのに伴って、新パリティの生成、およびパリティレコードの更新が必要とする。ｍ個のデータレコードとパリティレコードから構成されるパリティグループの新パリティ生成方法としては、（ａ）パリティグループのｍ個のデータレコードすべての排他的論理和により算出する方法の他に、排他的論理和の性質を利用して、（ｂ）パリティグループのいくつかの更新データレコードの旧データと新データとパリティレコードの旧パリティとの排他的論理和から算出する方法がある。通常、（ｂ）の方法の方が、新パリティの生成に必要な情報を揃えるためのオーバヘッドが小さいため、（ｂ）の方法を前提に以下の説明を行う。パリティ生成に必要な情報を揃えるための転送処理として、（ｂ）に示した値を得る処理を実行すると、パリティグループに更新データレコードが一つだけの場合でも、更新データレコードの旧データ、および対応する旧パリティをディスク装置から読み込むために２回、新データと新パリティをディスク装置に書き込むために２回、計４回の記憶制御装置とディスク装置間のデータ転送が生じ、キャッシュメモリの割り当てとディスク装置間のバス確保などのオーバヘッドが発生する。さらに、そろえた情報を元に新パリティを生成するオーバヘッドも生じる。

【００１０】特開平５－４６３２４号に開示された技術は、ディスク装置に新パリティを作成する機能を持たせることで、記憶制御装置の負荷をある程度分散させることができる。また、データ転送は、更新データレコードの新データをデータディスク装置とパリティディスク装置にブロードキャストするのに１回、更新データレコードの旧データを読み込んでパリティディスク装置に転送するのに１回、計２回で済むことになる。

【００１１】ただ、この技術では、ブロードキャストさ

れた新データをパリティディスク装置で受信しても、データディスク装置から対応する旧データを受信するまでは新パリティ生成を行なえない。このため、複数の更新データレコードの新データをディスク装置内のバッファメモリに保持することになるが、保持した複数の新データのスケジュールにはほとんど自由度がなく、あくまで旧データ到着順とせざるを得ない。仮に、平均シーク時間を短縮するべく、現ディスクヘッド位置から最短シークとなる新データから処理しようとするなら、旧データが未到着で処理可能でない新データとともに、旧データが到着して処理可能となった新データをも複数保持せねばならず、ディスク装置に搭載するバッファメモリを大容量のものにする必要が生じ、製造コストを押し上げることになる。

【００１２】一方、特開平７－２６１９４６号公報の技術では、データディスク装置に新旧データから中間値を作成する手段を設けることで、データ転送を、データディスク装置への新データ転送に１回、パリティディスク装置への新旧データの中間値転送に１回、計２回に抑えている。

【００１３】しかし、この技術では、新データと旧データの排他的論理和演算をデータディスク装置で、その結果生成した中間値と旧パリティの排他的論理和演算をパリティディスク装置で、それぞれ分割して実行するため、パリティディスク装置で三者の排他的論理和演算を一括実行する方法に比べて、各データを一時的に格納するバッファメモリへのロード／ストア実行回数が増加し、その結果、演算に要する総オーバヘッドは増大するものと考えられる。

【００１４】また、この技術では、データレコードへの更新に伴ってパリティレコードへの更新が発生し、互いに独立に動作しているデータディスク装置とパリティディスク装置へ同じ契機にアクセスを行なうことになる。このため、両ディスク装置の平均シーク時間の低減に対しては何ら考慮されていない。

【００１５】本発明の目的は、パリティ生成に要するオーバヘッドの軽減を図り、記憶装置サブシステム全体のスループット増大を図ることである。

【００１６】本発明の別の目的は、パリティ生成におけるデータレコードとパリティレコードの更新を非同期化し、各ディスク装置アクセスに伴う平均シーク時間を軽減し、スループット増大を図ることである。

【００１７】

【課題を解決するための手段】本発明では、記憶装置サブシステム全体のスループット増大を図るという目的を実現するために、以下の２つの観点で示した記憶制御方法を実現する記憶装置サブシステムを提供する。

【００１８】第一の観点では、記憶制御装置はデータ処理装置から更新され、キャッシュ上に蓄積したデータレコードのライトアフタ処理を行うために、同一パリティグループのデータレコードのうち、キャッシュ上に存在するものの新データを対応するデータディスク装置に送信する。データディスク装置では、受信した新データに対する旧データをリードした後、新データをディスクにライトし、新旧データを対応するパリティディスク装置へ転送する。一方、パリティディスク装置では、受信した新旧データに対応するパリティレコードの旧パリティをリードし、新パリティを生成してディスクにライトする。

【００１９】第二の観点でも、記憶制御装置は、第一の観点と同様に、同一パリティグループの複数のデータレコードの新データを各データディスク装置に転送し、データディスク装置では、対応する旧データをリード後、新データをライトする。第二の観点が第一の観点と異なるのは、ここでリードした旧データを一旦記憶制御装置に転送する点である。記憶制御装置では、ディスク装置への反映が完了し、旧データを読み込んだデータレコードをしばらくキャッシュ上に溜め置き、データディスク装置への反映とは非同期にパリティディスク装置への反映処理を実行する。

【００２０】これらの結果、パリティ生成のための排他的論理和演算はパリティディスク装置で一括して実行するので、演算の際のメモリロード／ストア回数を削減でき、オーバヘッドを軽減することができる。さらに、パリティレコードの更新をデータレコードの更新と非同期に実行可能となり、シーク時間を考慮したディスク装置の割り当てスケジューリングを行なうことで、各ディスク装置アクセス時の平均シーク時間を短縮することができる。

【００２１】以上の内容を適用すると、更新データのライトの際に、ディスク装置でのパリティ生成オーバヘッド削減、かつ平均シーク時間短縮によるディスク装置占有時間削減を実現することで、記憶装置サブシステム全体のスループットを増大させることが可能となる。

【００２２】

【発明の実施の形態】以下、本発明の実施例について説明する。なお、これにより本発明が限定されるものではない。

【００２３】本発明の実施例は、第１の実施例と第２の実施例がある。

【００２４】第１の実施例は、記憶制御装置から新データを受信したデータディスク装置で、旧データを記憶媒体から読み出し、受信した新データを記憶媒体に書き込み、新旧データを併せて、対応するパリティディスク装置に直接送信するものである。

【００２５】第２の実施例は、記憶制御装置から新データを受信したデータディスク装置で、旧データを記憶媒体から読み出し、受信した新データを記憶媒体に書き込み、旧データを記憶制御装置に送信し、記憶制御装置からパリティディスク装置へ新旧データを送信するもので

ある。

【００２６】まず、図１～図１１を参照して、第１の実施例を説明する。

【００２７】図１は、本発明の対象となる記憶装置サブシステムのブロック図である。

【００２８】記憶装置サブシステムは、データ処理装置１００と、記憶制御装置１０４と、１台以上のディスク装置１０５を接続してなっている。

【００２９】データ処理装置１００は、ＣＰＵ１０１と、主記憶１０２と、チャネル１０３とを有している。

【００３０】記憶制御装置１０４は、１つ以上のディレクタ１０６と、キャッシュメモリ１０７と、ディレクトリ１０８と、不揮発性メモリ１０９と、不揮発性メモリ管理情報１１０を有している。

【００３１】ディレクタ１０６は、チャネル１０３とディスク装置１０５の間のデータ転送、チャネル１０３とキャッシュメモリ１０７の間のデータ転送およびキャッシュメモリ１０７とディスク装置１０５の間のデータ転送を行なう。

【００３２】キャッシュメモリ１０７には、ディスク装置１０５の中のアクセス頻度の高いデータをロードしておく。このロード処理は、ディレクタ１０６が実行する。ロードデータの具体的な例は、ＣＰＵ１０１からアクセス対象となったデータレコードや、このデータレコードと物理ディスク装置１０５のディスク位置が近いデータレコードのデータ等である。また根、キャッシュメモリ１０７には、ＣＰＵ１０１から更新がかかり、ライトアフタで処理する新データを格納する。これらの新データは、キャッシュメモリ１０７に格納した段階でライト完了報告を出し、更新とは非同期にパリティ生成およびディスク装置１０５への書き込みを行う。

【００３３】ディレクトリ１０８は、キャッシュメモリ１０７の管理情報を格納する。

【００３４】不揮発性メモリ１０９は、不揮発の媒体であり、キャッシュメモリ１０７に格納した新データが物理ディスク装置１０５への反映が完了する前に消失しないよう、新データを格納し、キャッシュメモリ１０７と併せて二重管理する。

【００３５】不揮発性メモリ管理情報１１０も不揮発の媒体であり、不揮発性メモリ１０９の管理情報を格納する。

【００３６】ディスク装置１０５は、データレコードを記録するディスク１１０とデータを一時的に格納するバッファ１０９と記録されたデータを読み書きする装置から構成される。データを読み書きする装置は、排他的論理和演算により新パリティを生成する機能を有する。バッファ１１３は、複数のデータを格納することができる容量を持ち、必要があれば不揮発の媒体を用いる。

【００３７】各ディスク装置１０５間は互いに接続され、複数のディスク装置１０５間でのデータの送受信が可能である。

【００３８】図２は、本実施例におけるキャッシュメモリ１０７の構成である。

【００３９】キャッシュメモリ１０７は複数のセグメント２００で構成される。一つのセグメント２００には、１つのデータレコードもしくはパリティレコードを格納しうる。どのセグメント２００にどのレコードを割り当てるかについては、従来公知の方法を用いることができる。レコードが割り当てられていない空き状態のセグメント２００も存在しうる。

【００４０】図３は、ディレクトリ１０８の構造である。

【００４１】ディレクトリ１０８は、キャッシュ割り当て情報３００、セグメント管理情報３０１、空きセグメントキューポインタ３０２から構成される。なお、セグメント管理情報３０１は、セグメント２００の数だけ存在する。ただし、セグメント２００に静的に１対１対応するわけではない。以下、各情報を詳細に説明する。

【００４２】図４は、キャッシュ割り当て情報３００の構造である。データレコードやパリティレコードにセグメント２００が割り当てられ、キャッシュメモリ１０７上に格納されているかを表わす。レコード割り当てポインタ４００は、データレコードとパリティレコードに対応して設ける情報である。対応するデータレコードやパリティレコードがキャッシュ１０７上に格納されていれば、レコード割り当てポインタ４００は当該セグメント管理情報３０１をポイントする。格納されていなければヌル値となる。

【００４３】図５は、セグメント管理情報３０１の中で本発明に関係した情報を表わしたものである。ライト面セグメントポインタ５００について以下説明する。データレコードのセグメント管理情報３０１の場合、新パリティを生成していない新データを格納したセグメント２００をポイントする。パリティレコードのセグメント管理情報３０１の場合、キャッシュメモリ１０７上で新パリティを生成する際に、一時的に利用するセグメント２００をポイントする。以後、ライト面セグメントポインタ５０１でポイントするこれらのセグメント２００をライト面セグメントと呼ぶ。

【００４４】リード面セグメントポインタ５０１について以下説明する。データレコードのセグメント管理情報３０１の場合、新パリティを作成済みの値を格納したセグメント２００をポイントする。パリティレコードのセグメント管理情報３０１の場合、原則的に、本ポインタが示すセグメント２００にその値を格納する。以後、リード面セグメントポインタ５０１でポイントするこれらのセグメント２００をリード面セグメントと呼ぶ。

【００４５】ホストダーティフラグ５０２は、ライト面セグメントの値がデータ処理装置１００から書き込まれたままで、パリティ生成もディスク装置１０５への反映

も行なわれていないことを示す。物理ダーティフラグ５０２は、リード面セグメント値がパリティ生成が完了したが、まだディスク装置１０５に書き込まれていないことを示す。パリティ未反映フラグ５０２は、ライト面セグメントの値が、ディスク装置１０５へ反映されたが、パリティ生成がまだ行なわれていないことを示す。ライト面バリッドフラグ５０３はライト面セグメントに、リード面バリッドフラグ５０４はリード面セグメントに有効な値が入っていることを示す。使用中フラグ５０５は、そのセグメント管理情報３０１が割り当てられているデータまたはパリティレコードが何らかの処理で使用中であることを示し、処理間の排他に用いられる。レコード状態５０６は、個々のデータまたはパリティレコードの状態を示す。状態としては、「正常」の他に、ディスク装置１０５の障害等が原因で書き込みが出来ないことを示す「ライト不可」、読み込みが出来ないことを示す「リード不可」などがあり、複数の状態が同時に設定されることもある。空きポインタ５０７は、空き状態にあるセグメント管理情報３０１同士を結合する。

【００４６】図６は、空きセグメントキューの構造である。図に示すように、空き状態にあるセグメント２００の管理情報３０１を空きセグメントキューポインタ３０２に結合する。

【００４７】本実施例におけるキャッシュメモリ管理方法の説明の最後に、各セグメント２００に格納されるデータまたはパリティレコードの状態について述べる。

【００４８】図７は、セグメント２００に格納されるデータまたはパリティレコードの状態遷移の様子である。処理装置１００から更新がかかり、パリティ生成を行う前で、ライト面セグメントに新データを格納し、ライト面バリッドフラグ５０３とホストダーティフラグ５０２がオンの状態をホストダーティ状態７００と呼ぶ。ホストダーティ状態７００から、当該データレコードに対するパリティ生成が完了し、新データを格納したライト面セグメントをリード面セグメントに切り替え、リード面バリッドフラグ５０４と物理ダーティフラグをオンに、ホストダーティフラグ５０２をオフにした状態を物理ダーティ状態７０１と呼ぶ。キャッシュ１０７上で生成した新パリティも物理ダーティ状態７０１で保持する。また、ホストダーティ状態から、当該データレコードのディスク装置１０５より旧データを読んでリード面セグメントに格納し、新データをディスク装置１０５へ反映し、ホストダーティフラグ５０２をオフ、パリティ未反映フラグをオンにした状態をパリティ未反映状態７０２とする。さらに、物理ダーティ状態７０１のセグメント２００に対して、当該データまたはパリティレコードのディスク装置１０５への反映が完了し、物理ダーティフラグ５０２をオフにした状態、および、パリティ未反映状態７０２のセグメント２００に対して、当該データレコードのパリティ生成が完了し、パリティ未反映フラグ５０２をオフにした状態をクリーン状態７０３と呼ぶ。

【００４９】以下の説明では、あるセグメント２００をホストダーティ状態、物理ダーティ状態、パリティ未反映状態、クリーン状態にすることを、単にホストダーティ化、物理ダーティ化、パリティ未反映化、クリーン化と呼ぶ。また、あるデータ又はパリティセグメント２００の割り当てを解除して、セグメント２００およびセグメント管理情報３０１を空き状態にすることを、セグメント破棄と呼ぶ。セグメント２００破棄の際には、設定されているフラグ、状態、ポインタを全てクリアする。

【００５０】図８は、本実施例におけるディスク装置内バッファ１０９の構成である。バッファ１０９は複数のバッファセグメント８００で構成される。一つのバッファセグメント８００には１物理レコード分の値を格納でき、ディスク装置１０５と記憶制御装置１０４間でやりとりするデータまたはパリティレコードの値を格納する。リード処理時はディスク１１０からリードされた値が一時格納され、ライト処理時には記憶制御装置１０４から転送された書き込み値が一時格納される。

【００５１】次に本実施例での記憶制御装置１０４とディスク装置１０５の動作を説明する。

【００５２】図９は、本実施例における記憶制御装置１０４とディスク装置１０５の動作を詳細に表わした図である。

【００５３】まず、データ処理装置１００によりライトされたデータレコード９２０に対してセグメント２００を割り当て、新データを格納する。この段階で、ライト処理の完了をデータ処理装置１００に報告し、ホストダーティ状態の当該データレコード９２０をライト要求とは非同期にディスク装置１０５に反映する。

【００５４】この第１の実施例では、キャッシュ１０７上に保持されているホストダーティ状態のデータレコード９２０の中から、ライトアフタ処理対象となるものを選択し、対象データレコード９２０の新データを、当該データを格納するデータディスク装置１０５に送出する。次に、データディスク装置１０５では、受信した新データに対し、旧データのディスク１１０からの読みだしと新データのディスク１１０への書き込み、および対応するパリティを格納するパリティディスク装置１０５への新旧データの送出を行なう。以下、データディスク装置１０５でのこれら一連の処理をリード＆ライト処理と呼ぶ。新旧データを受信したパリティディスク装置１０５では、旧パリティのディスク１１０からの読み出しと新パリティ生成、および新パリティのディスク１１０への書き込みを行なう。以下、パリティディスク装置１０５でのこれら一連の処理をパリティ生成ライト処理と呼ぶ。最後に、要求したデータディスク装置からの処理完了報告を受けたライト転送処理９００が、当該更新データレコード９２０のセグメント２００をクリーン化、およびパリティレコードのセグメント２００を破棄す

る。

【００５５】以下では、ライトアフタ処理９００と、ディスク装置コマンド実行処理９１０について順に説明する。

【００５６】第一に、ライトアフタ処理９００について説明する。

【００５７】記憶制御装置１０４からディスク装置１０５への処理要求は、処理指定情報と処理対象データから構成される。指定情報は以下のような内容である。

【００５８】指定情報１…要求処理種別を指定する。要求処理には、リード、ライト、リード＆ライト、パリティ生成ライトなどがある。本実施例において、対象更新データレコードの新データを送信する際には要求処理種別にはリード＆ライトを、データディスク装置１０５からパリティディスク装置１０５へ新旧データを送信する際にはパリティ生成ライトをそれぞれ設定する。

【００５９】指定情報２…ディスク装置での処理対象範囲を指定する。

【００６０】処理対象データには、ライト要求やリード＆ライト要求の場合には更新データの新データが該当する。

【００６１】ライトアフタ処理９００では、セグメント２００に格納されている更新データレコード９２０のリード＆ライト要求をデータディスク装置１０５に送信し、データレコードとパリティレコードの更新完了を待って、データレコードに割り当てたセグメント２００のクリーン化などの処理を行なう。

【００６２】図１０はライトアフタ処理９００の処理フロー図である。

【００６３】ステップ１０００では、セグメント２００に格納され、キャッシュ１０７上に蓄積されたホストダーティ状態のデータレコードの中から、処理対象を選び出す。このとき、同一パリティグループに属する複数の更新データレコードが存在する場合は、同時に処理対象とする。

【００６４】ステップ１００１では、ステップ１０００で選出した各更新データレコード９２０に対応するパリティレコードにセグメント２００を割り当て、当該セグメント管理情報３０１の使用中フラグ５０５をオンにする。こうして、他処理による当該パリティグループのデータに対するパリティ生成を禁じ、排他を実現する。

【００６５】ステップ１００２では、ステップ１０００で選出した各更新データレコード１４２０のセグメント管理情報３０１の使用中フラグ５０５をオンにする。

【００６６】ステップ１００３では、処理対象の更新データレコード９２０の新データと処理指定情報からなるリード＆ライト要求を各データディスク装置１０５に送出し、データディスク装置１０５におけるリード＆ライト処理、およびパリティディスク装置１０５におけるパリティ生成ライト処理を行なわせる。

【００６７】ステップ１００４では、データディスク装置１０５とパリティディスク装置１０５に対する更新の完了した更新データレコード９２０のセグメント２００をクリーン化し、使用中フラグ５０５をオフする。ここで、セグメント２００を破棄ではなく、クリーン化するのは、以降のアクセスによる当該データレコードへのキャッシュヒットを期待するためである。

【００６８】ステップ１００５では、当該更新データレコード９２０のライト処理のために割り当てたパリティセグメント２００を破棄する。

【００６９】第二に、ディスク装置コマンド実行処理９１０について説明する。

【００７０】ライト転送処理９００により転送されたリード＆ライト要求を受信したデータディスク装置１０５では、当該データレコードの旧データリードと新データライトを実行後、パリティディスク装置１０５へ新旧データと共にパリティ生成ライト要求を送信する。パリティ生成ライト要求を受信したパリティディスク装置１０５では、旧パリティリードと新パリティ生成、および新パリティライトを行なう。

【００７１】図１１はディスク装置コマンド実行処理９１０の本発明に関係した部分のみを抜粋した処理フロー図である。

【００７２】ステップ１１００では、受信した処理指定情報の要求処理種別を調べる。処理種別がリード＆ライトの場合にはステップ１１０１に遷移し、パリティ生成ライトの場合にはステップ１１０６にジャンプする。

【００７３】ステップ１１０１では、処理指定情報とともに送信されてきた対象データレコードの新データをディスク装置１０５内のバッファセグメント９３０に取り込む。

【００７４】ステップ１１０２では、指定情報２で指定されたディスク１１０上の格納位置から、対象データレコードの旧データをリードし、新データを格納したのとは別のバッファセグメント９３０に格納する。

【００７５】ステップ１１０３では、ディスク１１０上の格納位置に、バッファセグメント９３０に格納した対象データレコードの新データをライトする。

【００７６】ステップ１１０４では、各々バッファセグメント９３０に格納した新旧データと処理指定情報からなるパリティ生成ライト要求をパリティディスク装置１０５に送出する。

【００７７】ステップ１１０５では、要求処理が完了したことをライトアフタ処理９００に報告し、処理を終了する。

【００７８】ステップ１１０６では、パリティ生成ライト処理を行なうために、処理指定情報とともに送信されてきた新旧データをバッファセグメント９３０に各々取り込む。

【００７９】このとき、対象データレコードが複数存在

する場合には、データレコード毎にパリティ生成ライトを行なっても良いし、全データレコード分の新旧データを同時にバッファセグメント９３０に取り込み、新パリティを一括生成しても構わない。また、パリティ生成ライト要求の新旧データを複数のパリティレコード分バッファ内に１０９にため込み、シーク時間最短のものから対象にするなど、適当なスケジューリングを行なっても構わない。

【００８０】ステップ１１０７では、パリティ生成ライト処理を行なう。まず、ディスク１１０から新パリティ生成に必要な旧パリティをリードして、バッファセグメント９３０に格納する。次にバッファ１０９上の新データと旧データ、および旧パリティから新パリティを生成し、また別のバッファセグメント９３０に格納する。最後に、ディスク１１０上の先ほどリードしたパリティ格納位置に、新パリティをライトする。

【００８１】次に、図１２～図１５を参照して、第２の実施例を説明する。

【００８２】図１２は、第２の実施例における記憶制御装置１０４とディスク装置１０５の動作を詳細に表した図である。

【００８３】この第２の実施例では、新データのデータディスク装置１０５への反映が終った段階で、データレコード９２０をパリティ未反映化して、キャッシュ上に保持し、パリティディスク装置１０５への反映処理を別の契機に行なう。まず、ホストダーティ状態のデータレコード９２０の中から、データ反映処理対象となるものを選択し、当該データレコード９２０の新データを、当該データレコードを格納するデータディスク装置１０５に送出する。次に、データディスク装置１０５では、受信した新データに対し、旧データのディスク１１０からの読みだしと新データのディスク１１０への書き込みを行ない、記憶制御装置１０４に旧データを送出する。以下、データディスク装置１０５での、これら一連の処理も第１の実施例と同様に、リード＆ライト処理と呼ぶ。記憶制御装置１０４では、旧データを受信したデータレコード９２０をパリティ未反映化して、当該データレコード９２０に対する処理を一旦打ち切る。キャッシュ１０７上に保留されたデータレコード９２０に対しては、この後、別の契機でパリティ反映処理を実行し、完了後、セグメント２００をクリーン化する。

【００８４】以下では、データ反映処理１２００と、パリティ反映処理１２１０と、ディスク装置コマンド実行処理１２２０について順に説明する。

【００８５】第一に、データ反映処理１２００について説明する。

【００８６】データ反映処理１２００では、更新データレコード９２０のリード＆ライト要求をデータディスク装置１０５に送信し、データレコードの更新完了を待って、受信した旧データをセグメント２００に格納し、データレコード９２０のパリティ未反映化などの処理を行なう。

【００８７】図１３はデータ反映処理１２００の処理フロー図である。

【００８８】ステップ１３００では、セグメント２００に格納され、キャッシュ１０７上に蓄積されたホストダーティ状態のデータレコード９３０の中から、処理対象を選び出す。

【００８９】ステップ１３０１では、ステップ１３０１で選出した更新データレコード９２０のセグメント管理情報３０１の使用中フラグ５０５をオンにする。

【００９０】ステップ１３０２では、処理対象の更新データレコード９２０の新データと処理指定情報からなるリード＆ライト要求をデータディスク装置１０５に送出し、リード＆ライト処理を行なわせる。

【００９１】ステップ１３０３では、データディスク装置１０５に対する反映の完了した更新データレコード９２０をパリティ未反映化し、使用中フラグをオフする。

【００９２】第二に、パリティ反映処理１２１０について説明する。

【００９３】パリティ反映処理１２１０では、データ反映処理１２００によってデータディスク装置１０５への反映処理が完了して、キャッシュ１０７上にパリティ未反映状態で蓄積されているデータレコード９２０のパリティ生成ライト要求をパリティディスク装置１０５に送信し、パリティレコードの更新完了を待って、データレコード９２０のクリーン化などの処理を行なう。

【００９４】図１４はパリティ反映処理１２１０の処理フロー図である。

【００９５】ステップ１４００では、セグメント２００に格納され、キャッシュ１０７上に蓄積されたパリティ未反映状態のデータレコード９３０の中から、処理対象を選び出す。

【００９６】ステップ１４０１では、ステップ１４０１で選出した更新データレコード９２０のセグメント管理情報３０１の使用中フラグ５０５をオンにする。

【００９７】ステップ１４０２では、処理対象の更新データレコード９２０の新旧データと処理指定情報からなるパリティ生成ライト要求をパリティディスク装置１０５に送出し、パリティ生成ライト処理を行なわせる。

【００９８】ステップ１４０３では、パリティディスク装置１０５に対する反映の完了した更新データレコード９２０をクリーン化し、使用中フラグをオフする。

【００９９】第三に、ディスク装置コマンド実行処理１２２０について説明する。

【０１００】データ反映処理１２００により転送されたリード＆ライト要求を受信したデータディスク装置１０５では、当該データレコードの旧データリードと新データライトを実行後、記憶制御装置１０４へ旧データを転送する。また、パリティ反映処理１２１０からパリティ

生成ライト要求を受信したパリティディスク装置１０５では、旧パリティリードと新パリティ生成、および新パリティライトを行なう。

【０１０１】図１５はディスク装置コマンド実行処理１２２０の本発明に関係した部分のみを抜粋した処理フロー図である。

【０１０２】ステップ１５００からステップ１５０７のうち、ステップ１５０４以外は図１１と全く同じであるので、説明は繰り返さない。

【０１０３】ステップ１５０４では、ステップ１５０２でディスク１１０から読み込み、バッファセグメント９３０に格納した旧データを記憶制御装置１０４に送出する。

【０１０４】

【発明の効果】本発明の記憶制御方法および記憶装置サブシステムによれば、データライト時の記憶制御装置の負荷軽減、かつ平均シーク／サーチ時間短縮を実現したパリティ生成ライトが可能となり、記憶制御装置のオーバヘッド削減の助長やドライブ占有時間削減などの効果により記憶装置サブシステム全体のスループットを増大させることができる。

【図面の簡単な説明】

【図１】本発明が対象とする記憶装置サブシステムのブロック図である。

【図２】本発明の実施例におけるキャッシュメモリの構成例示図である。

【図３】本発明の実施例におけるディレクトリの構成例示図である。

【図４】本発明の実施例におけるキャッシュ割り当て情報の構成例示図である。

【図５】本発明の実施例におけるセグメント管理情報の構成例示図である。

【図６】本発明の実施例における空きセグメントキューの構成例示図である。

【図７】本発明の実施例におけるセグメント状態遷移図である。

【図８】本発明の実施例におけるディスク装置内のバッファの構成例示図である。

【図９】本発明の第１の実施例において、キャッシュ上に保持された、データ処理装置からの新データをディスク装置へ書き込む際の、制御装置およびディスク装置の動作を示すブロック図である。

【図１０】本発明の第１の実施例において、更新データレコードのリード＆ライト要求を行うライトアフタ処理の処理フロー図である。

【図１１】本発明の第１の実施例において、記憶制御装置からの要求を処理するディスク装置コマンド実行処理の処理フロー図である。

【図１２】本発明の第２の実施例において、キャッシュ上に保持された、データ処理装置からの新データをディスク装置へ書き込む際の、制御装置およびディスク装置の動作を示すブロック図である。

【図１３】本発明の第２の実施例において、更新データレコードのリード＆ライト要求を行うデータ反映処理の処理フロー図である。

【図１４】本発明の第２の実施例において、パリティ未反映状態のデータレコードのパリティ生成ライト要求を行うパリティ反映処理の処理フロー図である。

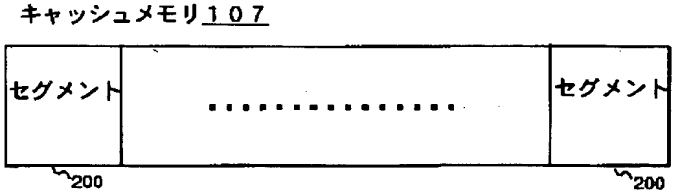
【図１５】本発明の第２の実施例において、記憶制御装置からの要求を処理するディスク装置コマンド実行処理の処理フロー図である。

【符号の説明】

１００…データ処理装置、　１０１…ＣＰＵ、　１０２…主記憶、１０３…チャネル、　１０４…記憶制御装置、１０５…ディスク装置、１０６…ディレクタ、　１０７…キャッシュメモリ、１０８…キャッシュディレクトリ、１０９…バッファ、１１０…ディスク、１１１…ＮＶＳ、　１１２…ＮＶＳ管理情報、２００…セグメント、３０１…セグメント管理情報、８００…バッファセグメント、９００…第１の実施例のライトアフタ転送処理部、９１０…第１の実施例のディスク装置コマンド実行処理部、９２０…キャッシュ上のデータセグメント、９３０…バッファに格納されたライトアフタ要求、１２００…第２の実施例のデータ反映処理部、１２１０…第２の実施例のパリティ反映処理部、１２２０…第２の実施例のディスク装置コマンド実行処理部。

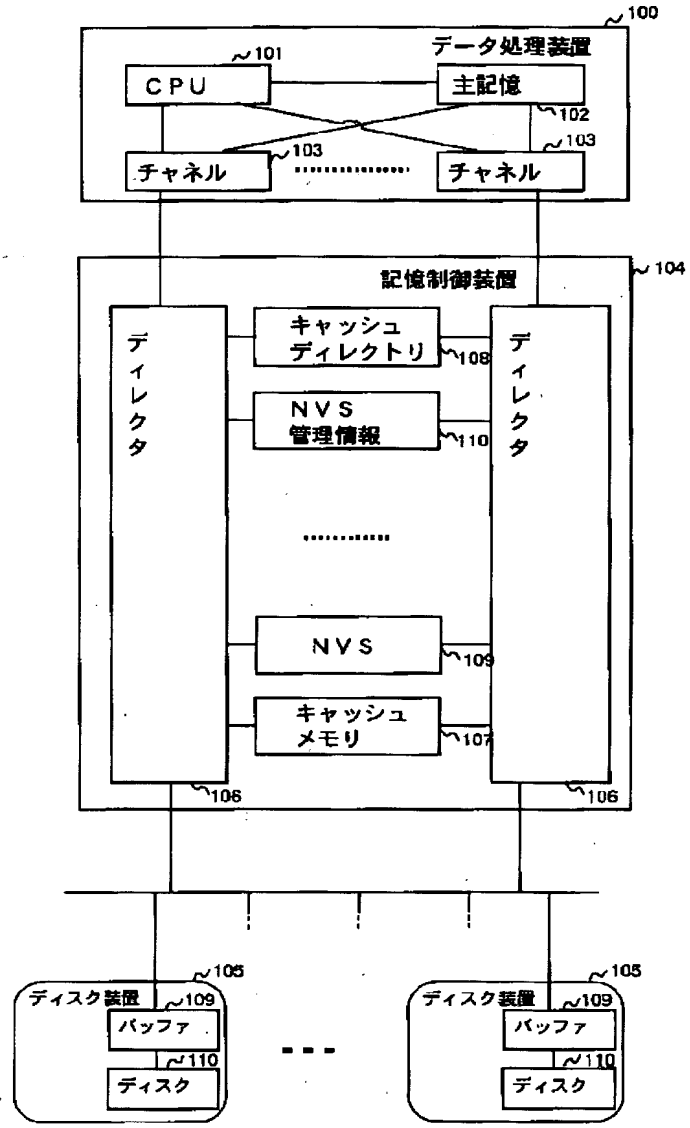
【図２】

図2



【図１】

図1



【図３】

図3

キャッシュディレクトリ１０８

キャッシュ割り当て情報	300
セグメント管理情報	301
：	
セグメント管理情報	301
空きセグメントキューポインタ	302

【図５】

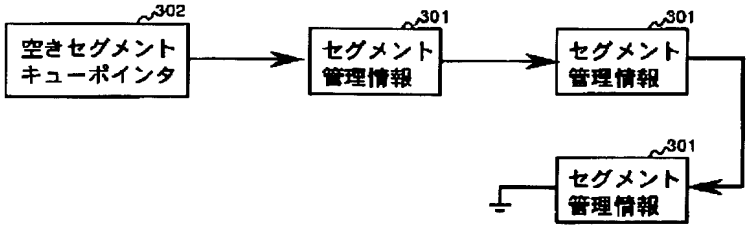
図5

セグメント管理情報３０１

ライト面セグメントポインタ	500
リード面セグメントポインタ	501
ホストダーティフラグ	502
物理ダーティフラグ	503
パリティ未反映フラグ	504
ライト面バリッドフラグ	505
リード面バリッドフラグ	506
使用中フラグ	507
レコード状態	508
キューポインタ	509

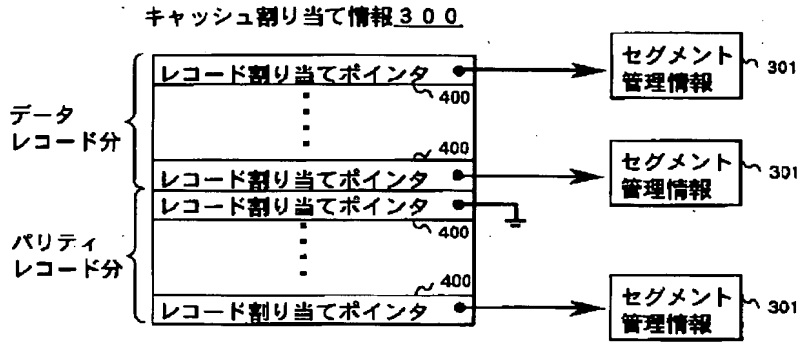
【図６】

図6



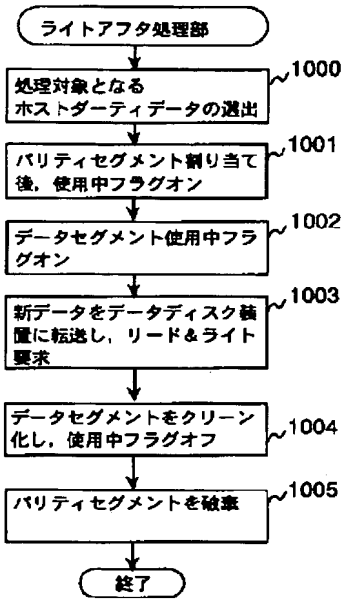
【図４】

図4



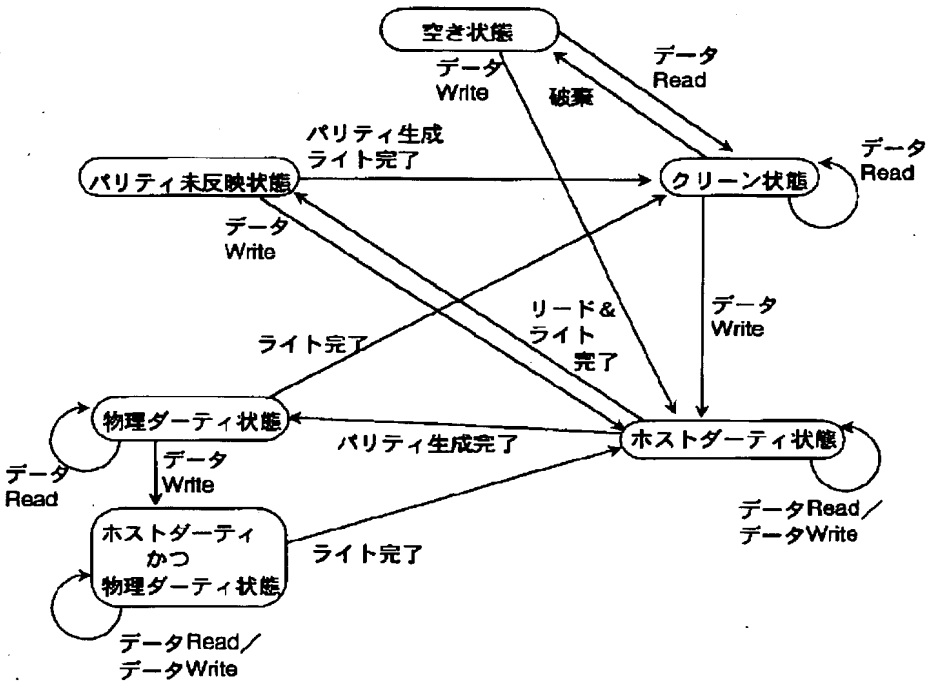
【図１０】

図１０



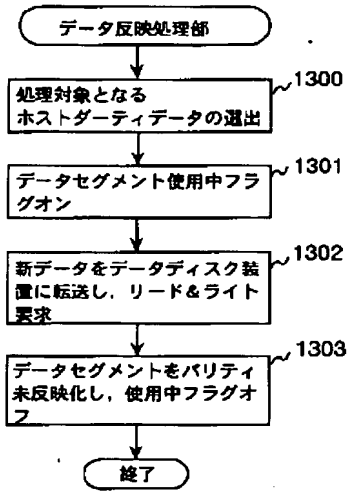
【図７】

図7



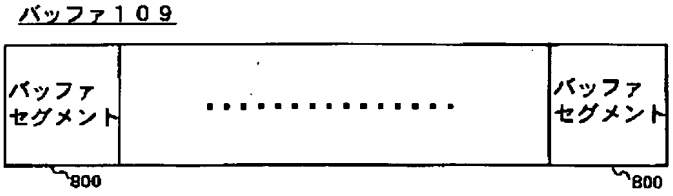
【図１３】

図１３



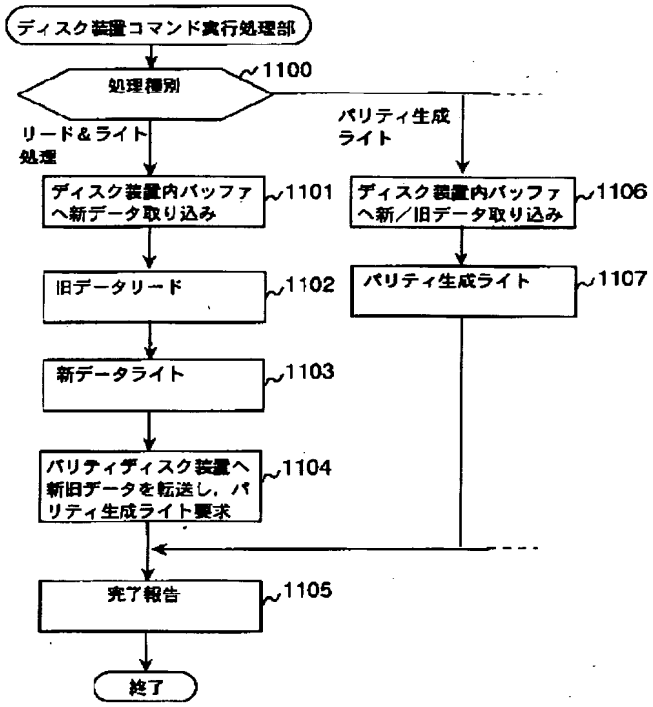
【図８】

図8



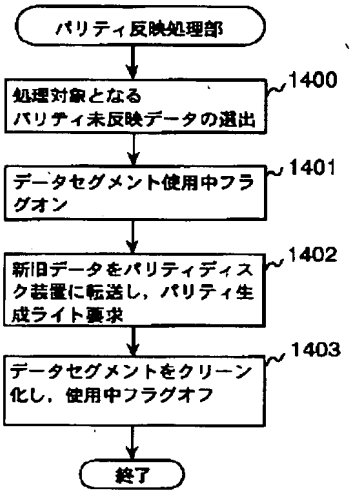
【図１１】

図１１



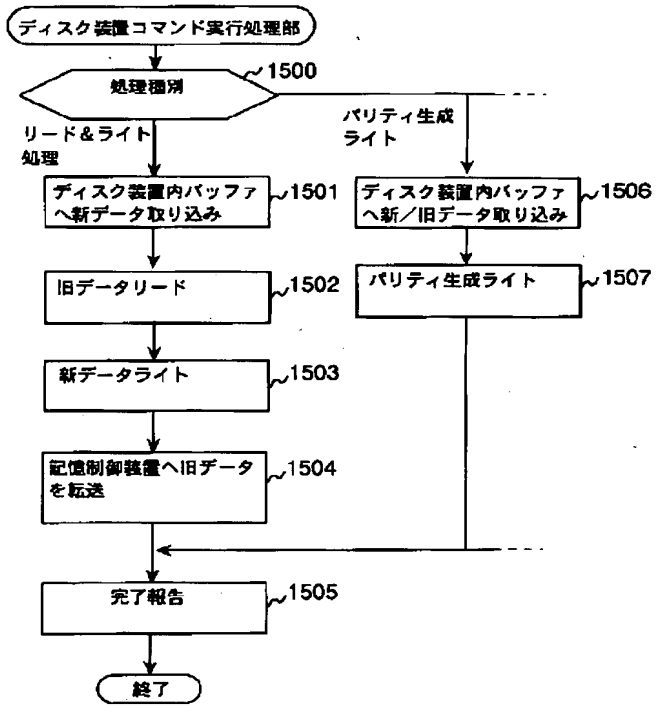
【図１４】

図１４



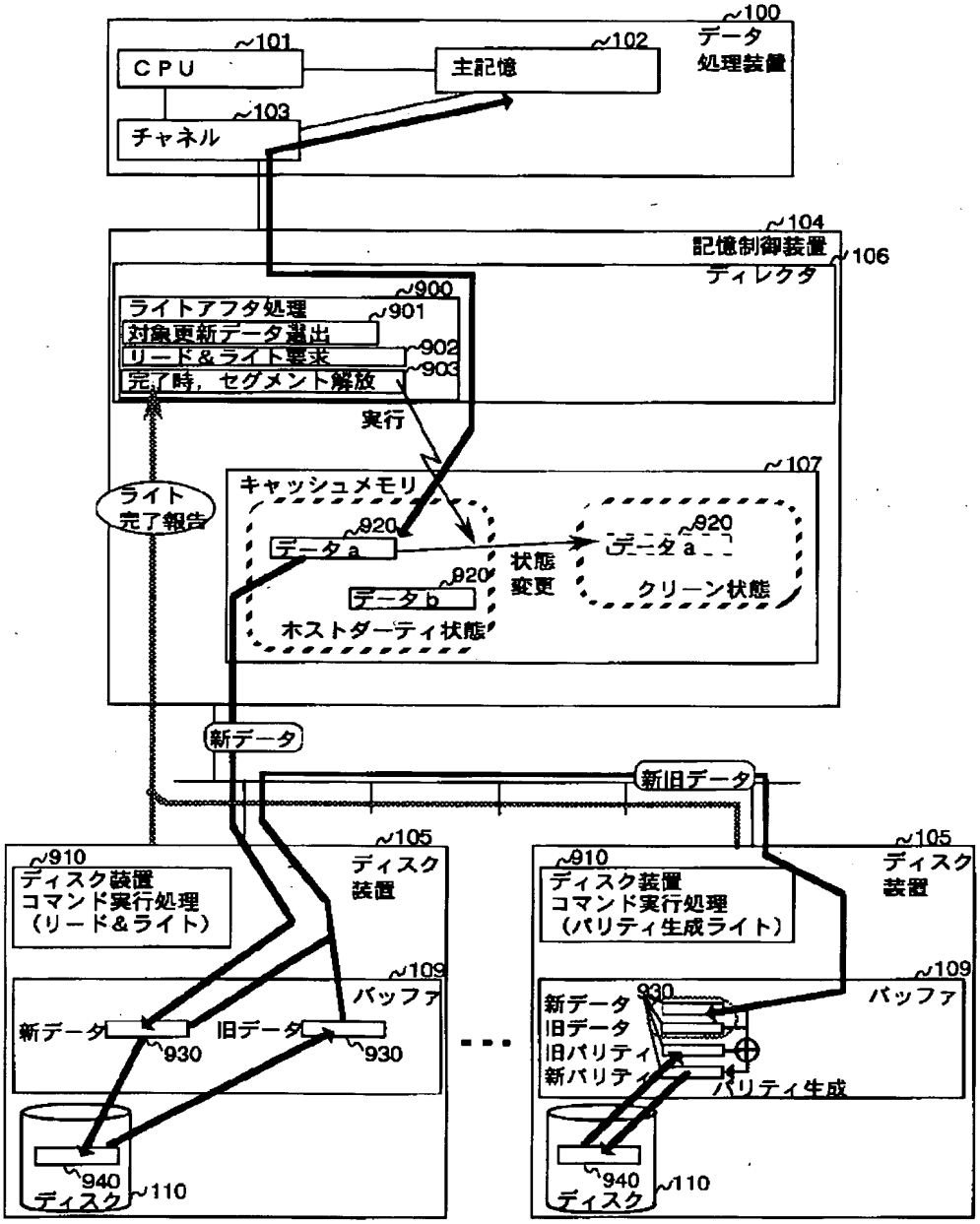
【図１５】

図１５



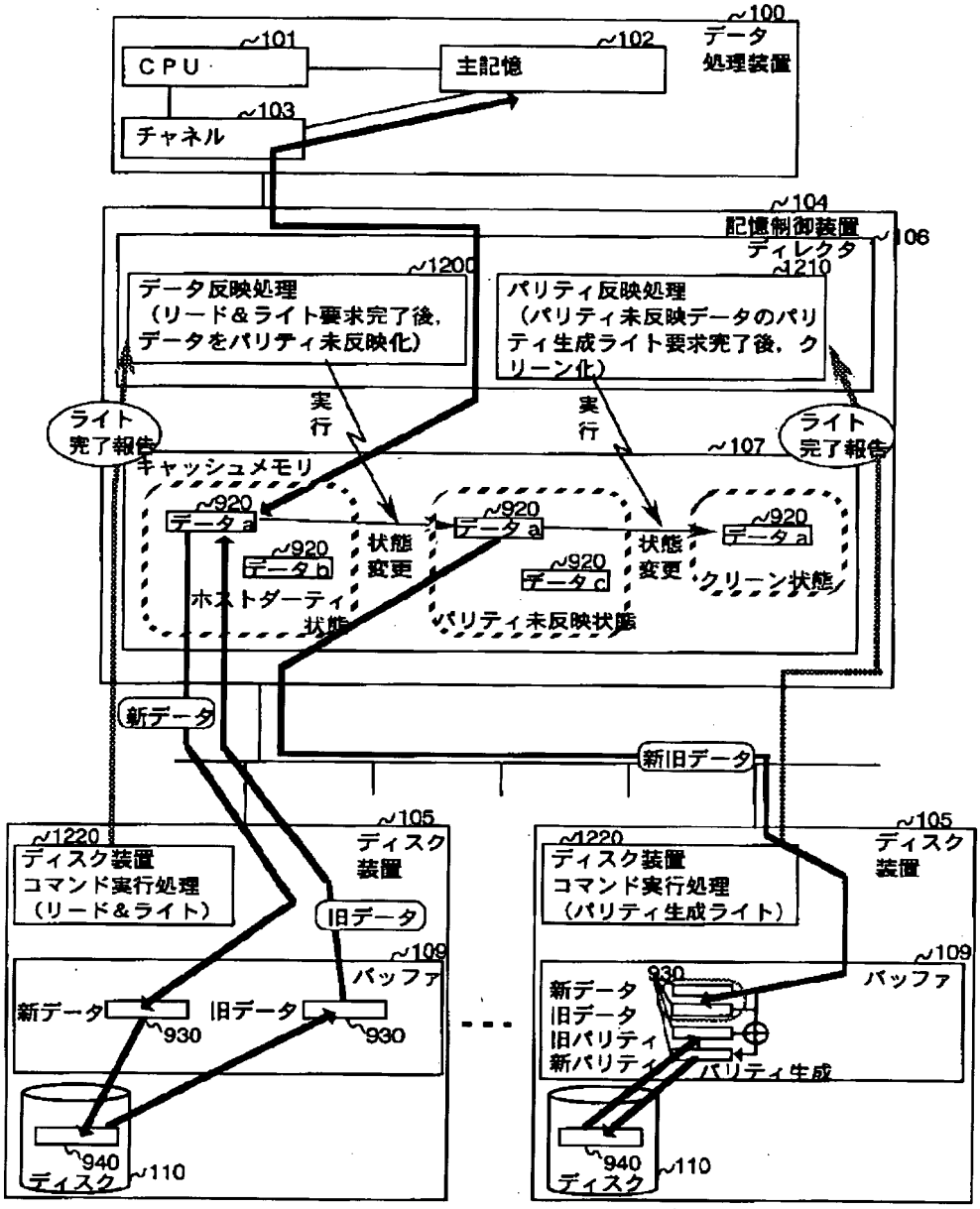
【図９】

図9


100
データ処理装置
101
CPU
102
主記憶
103
チャネル
104
記憶制御装置
106
ディレクタ
900
ライトアフタ処理
901
対象更新データ選出
902
リード＆ライト要求
903
完了時，セグメント解放
実行
107
キャッシュメモリ
ライト完了報告
920
データa
920
データa
状態変更
920
データb
クリーン状態
ホストダーティ状態
新データ
新旧データ
105
ディスク装置
910
ディスク装置コマンド実行処理（リード＆ライト）
109
バッファ
新データ
旧データ
930
930
940
ディスク
110
105
ディスク装置
910
ディスク装置コマンド実行処理（パリティ生成ライト）
109
バッファ
新データ
旧データ
旧パリティ
新パリティ
930
パリティ生成
940
ディスク
110

【図１２】

図１２


100
データ処理装置
101
CPU
102
主記憶
103
チャネル
104
記憶制御装置
106
ディレクタ
1210
1200
データ反映処理
（リード＆ライト要求完了後，データをパリティ未反映化）
パリティ反映処理
（パリティ未反映データのパリティ生成ライト要求完了後，クリーン化）
実行
実行
ライト完了報告
ライト完了報告
107
キャッシュメモリ
920
データa
920
データb
920
データa
920
データc
920
データa
状態変更
状態変更
ホストダーティ状態
パリティ未反映状態
クリーン状態
新データ
新旧データ
105
ディスク装置
1220
ディスク装置コマンド実行処理（リード＆ライト）
旧データ
109
バッファ
新データ
旧データ
930
930
940
ディスク
110
105
ディスク装置
1220
ディスク装置コマンド実行処理（パリティ生成ライト）
109
バッファ
930
新データ
旧データ
旧パリティ
新パリティ
パリティ生成
940
ディスク
110